ベイズ統計を用いた文書ファイルの自動分析手法

A method to analyze a set of documents besed on Bayesian statistics

後藤 正幸† 伊藤 潤‡ 石田 崇‡ 酒井 哲也* 平澤 茂一**

‡武蔵工業大学 環境情報学部 †早稲田大学 大学院 理工学研究科
*(株) 東芝 研究開発センター **早稲田大学 理工学部 経営システム工学科

要旨：近年，インターネット上には膨大な文書データが溢れており，すでに人間が全てを読んでそれらを体系化したり，必要な情報を分類整理することが困難になっている．そのため，従来からの情報検索技術の研究がさかんに行われており，目的に合致する情報を効率的に発見する試みが実用化されている．本稿では，情報検索技術の一手法である潜在意味的インデックシング (PLSI) という手法を用いて，文書データから知識発見を行う方法について述べる．そして文書データの解析では設定するモデルがデータ量に比べて相対的に複雑である点に着目し，ベイズ統計に基づく手法を提案する．さらにシミュレーション実験と応用実験を通じて，その性能を検討する．

Abstract: In this paper, we consider the Bayesian approach for representation of a set of documents. In the field of representation of a set of documents, many previous models, such as the latent semantic analysis (LSA), the probabilistic latent semantic analysis (PLSA), the Semantic Aggregate Model (SAM), the Bayesian Latent Semantic Analysis (BLSA), and so on, were proposed. In this paper, we formulate the Bayes optimal solutions for estimation of parameters. From the simulation experiments, we can show the effectiveness of the proposal. Moreover, we apply the proposal to analyze the questionnaires with free forms.

1 はじめに

近年，インターネット上には膨大な文書データが溢れており，すでに人間が全てを読んでそれらを体系化したり，必要な情報を分類整理することが困難になっている．そのため，従来からの情報検索技術の研究がさかんに行われており，目的に合致する情報を効率的に発見する試みが実用化されている．しかしながら，このような情報検索技術は，単に膨大なデータから効率的に必要な情報を探し出すための手法に留まらず，大量の文書データから効率的に知識発見を行うテキストマイニングの技術として応用が可能である．近年では，Web を用いて多くの自由記述式のアンケート結果や意見の集合を集めることが可能になっており，このような大量の文書集合を上手に解析し，知識を発見する技術は多くの恩恵をもたらすと考えられる．

本稿では，情報検索技術の一手法である潜在意味的インデックシング (PLSI) という手法を用いて，文書データから知識発見を行う方法について述べる．PLSI は，確率モデルを利用して，文書－単語行列を低次元に圧縮する方法であるが，これは裏を返せば，大量の文書データから特徴を凝縮して取り出すことを意味している．本稿では，これを知識発見に応用する方法について述べる．さらに，文書データの解析では設定するモデルがデータ量に比べて相対的に複雑である点に着目し，ベイズ統計に基づく手法を提案する．ベイズ統計は，事前分布を設定することにより，比較的データ数が少ないときに推定精度が高い方法であり，精度の面から有効性を示す．

さらにシミュレーション実験と応用実験を通じて，その性能を検討する．具体的には，授業改善を目的とした学生の自由記述式アンケートの分析に提案した方法を適用し，その性能について述べ，今後の可能性について吟味する．本稿で提案した結果は，消費者アンケートの分析などに応用可能であると考えられる．

2 潜在意味モデル

本節では，まず PLSI の土台となった LSI という手法を示し，その上で PLSI を示したあと，これらを文書クラスタリングに応用する方法について述べる．さらに，従来の PLSI を用いた文書クラスタリング手法の改善点について述べ，手法として改善すべきポイントを明確にする．

2.1 文書－単語行列

LSI や PLSI では，各文書ファイルを生成するメカニズムには，背後に観測できない隠れ属性 $c \in \mathcal{C} = \{c_1, c_2, \cdots, c_k\}$ があるものと仮定する．例えば，新聞記事でも，社会面と政治面，スポーツ面では自ずとその文書ファイルの性質は変わるであろう．このように文書データの構造を変えるカテゴリのようなものが，通常の文書データにも存在するものと仮定するのである．各文書ファイルは，切り出された単語 $w \in \mathcal{W} = \{w_1, w_2, \cdots, w_d\}$ によって特徴付けられるものとする．ここでは，分析対象となる文書ファイルは n 個存在し，これを $d \in \mathcal{X} = \{d_1, d_2, \cdots, d_n\}$ と表す．これらをまとめた行列

$$X = (d_1, d_2, \cdots, d_n)^T \tag{1}$$

を文書－単語行列と呼ぶ．ただし，T は転置を表し，各文書 d_i は

$$d_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,d})^T$$

という d 次元ベクトルである．$x_{i,j}$ は文書 d_i に単語 j が含まれるか否か，あるいはその出現回数等の情報を表す数である．

2.2 LSI

S. Deerwester らは，意味的情報検索のモデルとして LSI (Latent Semantic Indexing) を提案した [3]．LSI では，単語文書行列 X を特異値分解 (SVD) によって

$$X = U\Sigma V^T \tag{2}$$

と分解する．このうち，主成分の大きい方から k 個を用いて

$$\hat{X} = U_k \Sigma_k V_k^T \tag{3}$$

とすることにより，X を k 次元の潜在意味空間に圧縮することでノイズの除去を行う．これは，行列 X と $\hat{X}$ の 2 乗誤差を最小にする圧縮となっている．

しかし，LSI による情報検索においては単語文書行列 X に idf 値などで ad-hoc な重み付けが必要であるなど，いくつかの問題がある．

2.3 PLSI

一方，T. Hofmann によって提案された PLSI (Probabilistic Latent Semantic Indexing)[2] は，LSI と同様の圧縮を確率モデルに基づいて行う手法である．

PLSI では，意味的な隠れ属性 c_l $(l = 1, 2, \ldots, k)$ のもとで，文書 d_i $(i = 1, 2, \ldots, n)$ と単語 w_j $(j = 1, 2, \ldots, d)$ の生起は独立であると考え，d_i と w_j の同時確率 $P(d_i, w_j)$ を

$$P(d_i, w_j) = \sum_k P(d_i|c_l)P(w_j|c_l)P(c_l), \tag{4}$$

のように表す．ここで，文書 d_i における単語 w_j の実際の出現回数を $n(d_i, w_j)$ とすると，データの対数尤度

$$L = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j), \tag{5}$$

を最大にする $P(c_l)$，$P(d_i|c_l)$，$P(w_j|c_l)$ を，EM アルゴリズムで以下の式を計算することにより最尤推定する．

E-step

$$P(c_l|d_i, w_j) = \frac{P(c_l)P(d_i|c_l)P(w_j|c_l)}{\sum_{l'} P(c_{l'})P(d_i|c_{l'})P(w_j|c_{l'})} \tag{6}$$

M-step

$$P(w_j|c_l) = \frac{\sum_i n(d_i, w_j)P(c_l|d_i, w_j)}{\sum_{i,j'} n(d_i, w_{j'})P(c_l|d_i, w_{j'})} \tag{7}$$

$$P(d_i|c_l) = \frac{\sum_j n(d_i, w_j)P(c_l|d_i, w_j)}{\sum_{i',j} n(d_{i'}, w_j)P(c_l|d_{i'}, w_j)} \tag{8}$$

$$P(c_l) = \frac{\sum_{i,j} n(d_i, w_j)P(c_l|d_i, w_j)}{\sum_{i,j} n(d_i, w_j)} \tag{9}$$

(6)〜(9) 式の計算は，実際には，過学習を避けるため Tempered EM を用いている [2]．

2.4 PLSI を用いた文書クラスタリング

PLSI の隠れ属性 c_l は，ひとつの概念を表していると捉えることができる．S.Hirasawa and W.Chu[7] は，c_l を用いて以下のように文書集合を S 個のクラスタにクラスタリングする手法を提案している．

[アルゴリズム] [7]

1. $k = S$ として，EM アルゴリズムにより $P(c_l)$，$P(d_i|c_l)$，$P(w_j|c_l)$ を求める．
2. 各文書 d_i を，$\max_l P(c_l|d_i) = P(c_{\hat{l}}|d_i)$ となる $c_{\hat{l}}$ に割り振る．
3. 各 c_l に割り振られた文書集合をそれぞれ S 個のクラスタとする．

さらに，伊藤ら [8] は，PLSI が初期値に依存する EM アルゴリズムを用いた手法であることに着目し，予め初期値として代表元を与える手法を提案し，自由記述式アンケート結果に適用してその有効性を示している．これにより，分類する目的に応じて隠れ属性の代表元を生成し，これを積極的に分析に利用することができる．

[初期値の与え方] [8]

分類する目的に応じて隠れ属性 c_l の代表元 $\boldsymbol{d}_l$ を S 個作成（または選択）する．代表元 $\boldsymbol{d}_l$ は，$\boldsymbol{d}_l = (t_{l,1}, t_{l,2}, \ldots, t_{l,d})$ で表されるものとする．ここで，$\sum_j P(w_j|c_l) = 0$ となる c_l があると (6) 式が計算できないため，以下のように補正した初期値を用いる．

$$P(w_j|c_l) = \frac{t_{l,j} + \alpha}{\sum_{j'}(f_{l,j'} + \alpha)}, \tag{10}$$

$$P(d_i|c_l) = 1/n, \tag{11}$$

$$P(c_l) = 1/k. \tag{12}$$

α は正の値をとるパラメータである．

[アルゴリズム] [8]

1. S 個の代表元を作成し，$k = S$ として初期値を設定する．
2. EM アルゴリズムにより，$P(c_l)$，$P(d_i|c_l)$，$P(w_j|c_l)$ を求める．
3. 各文書 d_i を，$\max_l P(c_l|d_i) = P(c_{\hat{l}}|d_i)$ となる $c_{\hat{l}}$ に分類する．

2.5 問題と本研究への展開

従来の手法は，文書データから得られた確率モデルを元に，各文書 d_i に対して $\max_l P(c_l|d_i)$ を計算し，その最大値を与える c_l を用いてクラスタリングを行う手法であった．しかしながら，$P(c_l|d_i)$ は c_l 上の確率分布であるので，その最大値（モード）のみを用いてクラスタリングを行うと情報の損失がある．すなわち，一番高い確率を持つ隠れ属性 $c_{\hat{l}}$ 以外の情報を全く用いていない．

そこで，確率分布間の距離を表す指標として

$$I(d_i, d_j) = \sum_{c_l} P(c_l) \log \frac{P(c_l|d_i)}{P(c_l|d_j)}, \tag{13}$$

を用い，K-means 法や階層クラスタリング手法を適用することにより，文書を自動分類することが可能である．

一方，$P(c_l|d_i)$ は EM アルゴリズムによって推定された推定量であることを考慮すると，このクラスタリング手法は推定精度によってその有効性が左右されることがわかる．すなわち，より推定精度の高い手法を構成することが重要であるといえる．本稿では，確率モデルの推定精度向上という点に焦点を絞り，主に小サンプルで推定精度の良好なベイズ統計に基づく手法を提案する．PLSI で想定しているモデルは，隠れ属性をもつ比較的複雑なモデルであり，文書データ数の増加と共に，隠れ属性数も増加されるのが一般である．一方，隠れ属性数が大きいということは，推定すべき確率モデルのパラメータが通常の統計モデルで考えているよりも遥かに多いことを示しており，そのためベイズ統計的手法が有効となるものと考えられる．

以下では，ベイズ統計に基づく手法を提案し，実験を通じてその性能を示す．

3 BPLSI の提案と計算法

BLSA(Bayesian Latent Semantic Analysis) というベイズ統計に基づく手法がすでに提案されている [13]．しかしながら，この方法は EM アルゴリズムによって推定を行う方法であり，本質的には PLSI と大きく変わりがない．本稿では，PLSI が想定している確率モデルを確率モデル族とみなし，ベイズ決定理論に基づいてベイズ最適な推定方法，およびクラスタリング方法を導く．

3.1 基本モデル

PLSI のモデルでは，それぞれの文書 d_i $(i \in \{1,2,\cdots,n\})$ は

$$P(d_i|\theta,\lambda)=\sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}P(x_{i,j}|\theta_{l,j},c_l^k), \tag{14}$$

という混合確率分布に従って生起するものとする．

もし $x_{i,j}\in\{0,1\}$ であれば，$P(d_i|\theta,\lambda)$ はパラメータ $0<\theta_{l,j}<1$, $\sum_l\lambda_l=1$ を用いて

$$P(d_i|\theta,\lambda)=\sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}(\theta_{l,j})^{x_{i,j}}(1-\theta_{l,j})^{1-x_{i,j}},$$

の形で表される．ただし，$\theta=(\theta_1,\theta_2,\cdots,\theta_k)$, $\theta_l=(\theta_{l,1},\theta_{l,2},\cdots\theta_{l,d})$, $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_k)$ である．

このもとで，尤度関数 $P(X|\theta,\lambda)$ は

$$P(X|\theta,\lambda)=\prod_{i=1}^{n}\left\{\sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}(\theta_{l,j})^{x_{i,j}}(1-\theta_{l,j})^{1-x_{i,j}}\right\}$$

$$=\sum_{l_1=1}^{k}\cdots\sum_{l_n=1}^{k}\left\{\prod_{i=1}^{n}\lambda_{l_i}\prod_{j=1}^{d}(\theta_{l_i,j})^{x_{i,j}}(1-\theta_{l_i,j})^{1-x_{i,j}}\right\}, \tag{15}$$

と与えられる．

3.2 ベイズ最適解の導出

まず，パラメータ空間上に事前分布 $f(\theta_{l,j}|c_l^k)$, $f(\lambda)$ を設定する．このとき，データ X が与えられたもとでの事後確率 $f(\theta,\lambda|X)$ は

$$f(\theta,\lambda|X)=\frac{P(X|\theta,\lambda)f(\theta)f(\lambda)}{\int_\lambda\int_\theta P(X|\theta,\lambda)f(\theta)f(\lambda)d\theta d\lambda},$$

で与えられる．パラメータ推定のための損失として二乗誤差を考えると，ベイズ最適な推定量は

$$\left(\tilde{\theta},\tilde{\lambda}\right)=\int_\theta\int_\lambda(\theta,\lambda)P(\theta,\lambda|X)d\theta d\lambda,$$

で与えられる．

3.3 ベイズ最適解の計算法

パラメータ $P(\theta,\lambda|X)$ を推定するために，λ に対してディレクレ分布

$$f(\lambda)=\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\cdots\lambda_k^{\alpha_k-1},$$

を仮定する．ただし，$\Gamma(\cdot)$ はガンマ関数である．θ に対しては，ベータ分布

$$f(\theta_{l,j}|c_l^k)=\frac{\Gamma(\beta_1^{l,j}+\beta_2^{l,j})}{\Gamma(\beta_1^{l,j})\Gamma(\beta_2^{l,j})}(\theta_{l,j})^{\beta_1^{l,j}-1}(1-\theta_{l,j})^{\beta_2^{l,j}-1},$$

を仮定する．もし，事前知識が何もないときは，$\alpha_l=1$, $\beta_1^{l,j}=\beta_2^{l,j}=1$ とすればよい．

ここで，$l=(l_1,l_2,\cdots,l_n)$ とし，$\mathcal{L}=\{(l_1,l_2,\cdots,l_n)|l_i\in\{1,2,\cdots,k\}\}$ と記述することにする．$l=l_i$ は i 番目の文書が l 番目の隠れ属性 c_l から出現することを意味する．$\mathbf{1}\{\cdot\}$ をインジケータ関数として $z(l|\boldsymbol{l})$ を

$$z(l|\boldsymbol{l})=\sum_{i=1}^{n}\mathbf{1}\{l=l_i\}, \tag{16}$$

で定義しよう．$z(l|\boldsymbol{l})$ は，$\boldsymbol{l}$ 内における l 番目の隠れ属性 c_l の出現回数を意味する．さらに，$x(j|l,\boldsymbol{l})$ を

$$x(j|l,\boldsymbol{l})=\sum_{i:l=l_i}x_{i,j}, \tag{17}$$

と定義する．

以上の設定のもとに，全文書データの尤度関数は

$$P(X|\theta,\lambda)=$$

$$\sum_{\boldsymbol{l}\in\mathcal{L}}\prod_{l=1}^{k}\left\{(\lambda_l)^{z(l|\boldsymbol{l})}\prod_{j=1}^{d}(\theta_{l,j})^{x(j|l,\boldsymbol{l})}(1-\theta_{l,j})^{z(l|\boldsymbol{l})-x(j|l,\boldsymbol{l})}\right\},$$

と与えられるので，パラメータ (θ,λ) の事後確率密度は

$$f(\theta,\lambda|X)=$$

$$\frac{1}{K}\sum_{\boldsymbol{l}\in\mathcal{L}}\left\{\left(\prod_{l'=1}^{k}(\lambda_{l'})^{z(l'|\boldsymbol{l})+\alpha_{l'}-1}\right)\prod_{l=1}^{k}\prod_{j=1}^{d}\frac{\Gamma(\beta_1^{l,j}+\beta_2^{l,j})}{\Gamma(\beta_1^{l,j})\Gamma(\beta_2^{l,j})}\right.$$

$$\left.\cdot(\theta_{l,j})^{x(j|l,\boldsymbol{l})+\beta_1^{l,j}-1}(1-\theta_{l,j})^{z(l'|\boldsymbol{l})-x(j|l,\boldsymbol{l})+\beta_2^{l,j}-1}\right\}.$$

のように与えられる．ただし，基準化定数 K は

$$K=\sum_{\boldsymbol{l}\in\mathcal{L}}\left\{\frac{\prod_{l'=1}^{k}\Gamma(z(l'|\boldsymbol{l})+\alpha_{l'}-1)}{\Gamma(n+\alpha_{l'})}\right.$$

$$\left.\prod_{l=1}^{k}\prod_{j=1}^{d}\frac{\prod_{i=0}^{x(j|l,\boldsymbol{l})-1}\left(i+\beta_1^{l,j}\right)\prod_{i=0}^{z(l|\boldsymbol{l})-x(j|l,\boldsymbol{l})-1}\left(i+\beta_2^{l,j}\right)}{\prod_{i=0}^{n-1}\left(i+\beta_1^{l,j}+\beta_1^{l,j}\right)}\right\}.$$

で与えられる．

さらに，$K_\lambda(\boldsymbol{l})$ と $K_{\theta_{l,j}}(\boldsymbol{l})$ を

$$K_\lambda(\boldsymbol{l})=\frac{\prod_{l'=1}^{k}\Gamma(z(l'|\boldsymbol{l})+\alpha_{l'}-1)}{\Gamma(n+\alpha_{l'})}$$

$$K_\theta(\boldsymbol{l})=\prod_{l=1}^{k}\prod_{j=1}^{d}\frac{\prod_{i=0}^{x(j|l,\boldsymbol{l})-1}\left(i+\beta_1^{l,j}\right)\prod_{i=0}^{z(l'|\boldsymbol{l})-x(j|l,\boldsymbol{l})-1}\left(i+\beta_2^{l,j}\right)}{\prod_{i=0}^{n-1}\left(i+\beta_1^{l,j}+\beta_1^{l,j}\right)}$$

のように定義すれば，K は

$$K=\sum_{\boldsymbol{l}\in\mathcal{L}}K_\lambda(\boldsymbol{l})K_\theta(\boldsymbol{l}).$$

のように書き下すことができる．

この事後確率密度の平均を取ることにより，$\theta_{l,j}$ の最適な推定量 $\tilde{\theta}_{l,j}$ は

$$\tilde{\theta}_{l,j}=\frac{1}{K}\sum_{\boldsymbol{l}\in\mathcal{L}}\left\{K_\lambda(\boldsymbol{l})K_\theta(\boldsymbol{l})\left(\frac{x(j|l,\boldsymbol{l})+\beta_1^{l,j}}{n+\beta_1^{l,j}+\beta_2^{l,j}}\right)\right\}. \tag{18}$$

で与えられ，λ の最適な推定量 $\tilde{\lambda}_l$ は

$$\tilde{\lambda}_l=$$

$$\frac{1}{K}\sum_{l\in\mathcal{L}}\left\{K_\lambda(l)K_\theta(l)\left(\frac{z(l|l)+\alpha_l}{n+\alpha_1+\alpha_2+\cdots+\alpha_k}\right)\right\}. \tag{19}$$

で与えられる．ただし，

$$K_\lambda=\sum_{l\in\mathcal{L}}\frac{\prod_{l=1}^{k}\Gamma(z(l|l)+\alpha_l)K_{\theta_{l,j}}(l)}{\Gamma(n+\alpha_1+\alpha_2+\cdots+\alpha_k)}.$$

である．どちらも，ラプラス推定量の重み付け和の形で与えられることがわかる．

以上により，全ての l に対してラプラス型推定量を構成して，それを l の事後確率で重み付け和をとったものが，ベイズ最適な推定量であることがわかる．

4　実験と考察

本稿では小規模な実験を通じて提案法の有効性を検証してみる．従来手法である EM アルゴリズムに基づく PLSI を比較対象とした．隠れ属性数 3，単語数 10 の基本モデルに対し，一様乱数を使って確率パラメータを設定する．そのもとで，データ数（文書数）n を増やしながら，それぞれの n に対してデータを 1000 回発生させて平均二乗誤差を計算した．

評価指標である平均二乗誤差は，文書の生起確率 $P(d_i|c_l)$ に対して，推定量 $\hat{P}(d_i|c_l)$ の平均二乗誤差

$$\sum_{l=1}^{k}P(c_l)\{P(d_i|c_l)-\hat{P}(d_i|c_l)\}^2$$

と，単語の生起確率 $P(w_i|c_l)$ に対して，推定量 $\hat{P}(w_i|c_l)$ の平均二乗誤差

$$\sum_{l=1}^{k}P(c_l)\sum_{i=1}^{d}\{P(w_i|c_l)-\hat{P}(w_i|c_l)\}^2$$

とする．シミュレーション実験の結果，データ数が極めて少ないときではあるが，提案法が平均二乗誤差の面で優れていることがわかる．

しかしながら，提案法ではデータ数 n と隠れ属性数 k に対し，確率計算の計算量が $O(k^n)$ となってしまい，大きな n に対して計算が困難である．代表元を用いて少ない計算量で有効な近似解を求める，積と和の構造を用いて計算をまとめることにより計算量を低減したアルゴリズムを構成するなど，大規模データへの適用を可能とすることが今後の課題であろう．具体的な応用については紙面の都合上割愛する．

5　まとめ

本稿では，文書データのクラスタリングに適用可能な PLSI のモデルを，ベイズ統計の枠組みで再構築した．これにより，データ数が少ない時の推定精度の有効性が確認された．

大規模な文書データに適用可能な実務的アルゴリズムを構築することが今後の課題である．

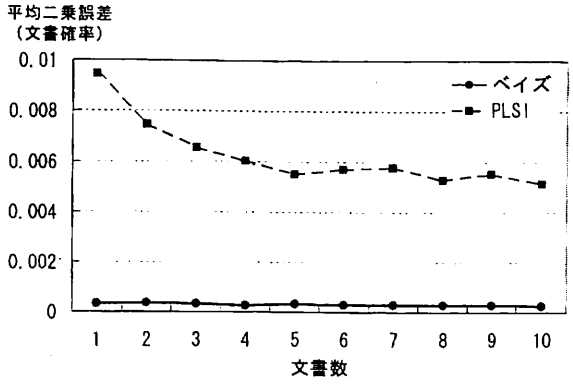

図 1: 平均二乗誤差（文書確率）

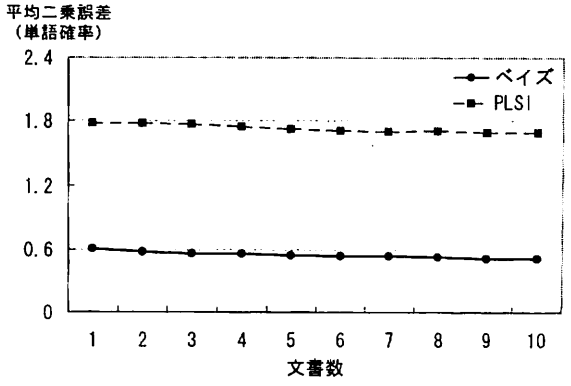

図 2: 平均二乗誤差（単語確率）

参考文献

[1] M. Hearst, "Untangling Text Data Mining," ACL '99 Proceedings, pp.3-10, 1999.

[2] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. of SIGIR'99, ACM Press, pp.50-57, 1999.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," J. of the Society for Information Science, 41, pp.391-407, 1990.

[4] 長坂悦敬，阿手雅博，"記述問題の自動評価を目指した教育支援情報システムによる Interactive Education，" 情報教育方法研究第 3 巻，第 1 号，pp.37-42, 2000 年.

[5] 酒井哲也，伊藤潤，後藤正幸，石田崇，平澤茂一，"情報検索技術を用いた効率的な授業アンケートの分析，" 経営情報学会 2003 年春季全国研究発表大会予稿集，pp.182-185，東京，2003 年 6 月.

[6] 後藤正幸，酒井哲也，伊藤潤，石田崇，平澤茂一，"選択式・記述式アンケートからの知識発見，" PC カンファレンス予稿集，鹿児島，2003 年 8 月.

[7] S. Hirasawa and W. W. Chu, "Knowledge acquisition from documents with both fixed and free formats," to appear in Proc. of 2003 IEEE Int. Conf. on System, Man, and Cybernetics, Washington DC, U.S.A., Oct. 2003.

[8] 伊藤潤，石田崇，後藤正幸，酒井哲也，平澤茂一，"PLSI を利用した文書からの知識発見，" 2003 年 FIT 論文集，vol.2. pp.83-84，江別，2003 年 9 月.

[9] 石田崇，伊藤潤，後藤正幸，酒井哲也，平澤茂一，"授業モデルとその検証，" 経営情報学会 2003 年秋季全国研究発表大会予稿集，函館，2003 年 11 月.

[10] 平澤茂一，石田崇，伊藤潤，後藤正幸，酒井哲也，"授業に関する選択式・記述式アンケートの分析，" 15 年度大学情報化全国大会，pp.144-145，東京，2003 年 9 月.

[11] 平澤茂一，コンピュータ工学，培風館，2001 年.

[12] 伊藤潤，石田崇，後藤正幸，平澤茂一，"文間の単語共起類似度を用いた重要文抽出手法，" 2002 年 FIT 論文集，pp.83-84，東京，2002 年 9 月.

[13] Nando de Freitas and Kobus Barnard, "Bayesian Latent Semantic Analysis," http://elib.cs.berkeley.edu/papers/clustering/bayesian/, 2000.